Research Article

Optimizing Location of Car-Sharing Stations Based on Potential Travel Demand and Present Operation Characteristics: The Case of Chengdu

Yu Cheng,1 Xu Chen,1 Xiaohua Ding,1 and Linting Zeng2

1 Shanghai Electric Vehicle Public Data Collecting Monitoring and Research Centre, 5F, No. 888, South Moyu Road, Jiading, Shanghai 201805, China
2 School of Environmental and Geographical Sciences, Shanghai Normal University, Shanghai, China

Correspondence should be addressed to Yu Cheng; chenyu@shevdc.org

Received 11 September 2018; Accepted 18 December 2018; Published 20 January 2019

Guest Editor: Ali Tizghadam

Copyright © 2019 Yu Cheng et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Car-sharing is becoming an increasingly popular travel mode in China and many companies invest plenty of money on that including vehicle enterprises and Internet companies. But most of them build car-sharing stations by their experience or randomly as long as there is parking space in the early development of their business. This results in many stations with low operational efficiency and causes capital loss. This study aims to use different data source with statistical models and machine learning algorithm to help car-sharing operator to choose the optimal location of new stations and adjust the location of existing stations. We select Chengdu where there are huge amounts of car-sharing travel demand and several large car-sharing operators as the research area and two main operators as the research objects. Chengdu is divided into 58724 squared grids each of which is 0.5 km × 0.5 km instead of focusing on the buffers generated by stations. We try to find a model to estimate a potential travel demand value for each small grid with three data sources: order data, population data, and Point of Interest (POI) data. This problem is transformed into a binary form and five different methods, Logistic Regression, Logistic Regression with LASSO, Naive Bayes, Linear Discriminant Analysis, and Quadratic Discriminant Analysis, are implemented. The optimal model, Logistic Regression with LASSO, is chosen to estimate the probability of existence of demand in all grids. With car-sharing order data from different operators, an existing order heat value is also computed for each grid. Then we analyze and classify all the grids into four groups. For different groups of grids, we give different suggestions on the optimal location of stations. This study focuses on a more competitive market and finds the influential factors on order number. Suggestions on the optimal location of stations are given in consideration of competitors. We hope that our research can help operators improve their business and make rational plans.

1. Introduction

Car-sharing is a mode that allows members to access a fleet of vehicles for short-term use without actual ownership. Members just need to reserve a vehicle online or by mobile app and then move to the parking lots and drive the car. They usually pay for this after travelling according to the travelling distance or/and time [1, 2]. This kind of new service allows people to avoid buying a car and spending time finding parking lots [3]. Car-sharing is becoming an increasingly important travelling mode in recent 5 years especially with rapid growth of electric vehicles since 2016. The advantages of car-sharing include reducing vehicle ownerships, reducing vehicle kilometres travelled, and reducing greenhouse gas emission [4–6].

So far, there are three main types of car-sharing mode: station-based, free-floating, and peer-to-peer car-sharing [7]. Station-based system requires the operators to hire parking space for stations and the vehicles are parcelled out among these stations. According to the operation mode, there are two types of station-based form: one-way car-sharing and round-trip car-sharing. One-way car-sharing allows the customers to return cars at any designated station wherever the trip started. In contrast, with round-trip car-sharing, cars should be returned to the station where the trip starts. The first form provides more flexibility but will cause the
problem of spatial imbalance over the stations. The operators have to consider relocation problem that leads to much cost. Free-floating system operates without station. Instead, the operators define an area where the customers can park cars at any parking lots [7, 8]. Free-floating car-sharing could also be treated as a special case of one-way car-sharing [9].

Our research will mainly focus on one-way station-based car-sharing system for which station siting is a big challenge and need to take quite a lot of factors into consideration. A good choice can bring high efficiency, large profit, low operation cost, and competitive advantages when competing with other operators [10, 11]. However, at the beginning of business, operators usually locate stations by their experience. For example, trading area, universities, or airports always bring great profit if there are car-sharing stations. But different cities have different features and attitudes towards car-sharing. Therefore, experience sometimes may be unreliable and relocations are necessarily introduced when the business is mature. What is more, some operators just choose locations randomly as long as there are parking spaces depending on their huge funds in order to seize the market ahead of other competitors, where relocations are even more necessary.

Researchers have done a lot of work on relocation or site selection problem by using various methods. Analytic hierarchy process (AHP) is the most popular method among multicriteria decision making methods [12]. One study uses AHP to solve the site selection problem for EVCARD which is a car-sharing operator in Shanghai. The researchers consider the potential users, potential travel demand, potential travel purposes, and distances from existing stations totally 15 factors as the decision criteria. But this method is based on candidate stations and expert scoring method that is subjective [11]. Mathematical and statistical models are also applied. Several studies use multilinear model and mixed-integer programming model to find the optimal location of stations [13–15]. Another paper introduces an intensity model that estimates the demand and an imbalance model that describes the difference between pick-up and drop-off to find the optimal location of stations of EVCARD in Shanghai, which involves usage intensity, usage imbalance, transportation information, and built environment. These two models utilize a combination of Elastic Net, the adaptive Least Absolute Shrinkage and LASSO [3].

The previous two studies about site location of EVCARD all focus on the market in Shanghai where there is no competitor. Therefore, it is much easier to make decision on the optimal site location. Our research will focus on a more complicated market: Chengdu where there are more than five car-sharing operators and each has its own advantages. In Chengdu, there are two main car-sharing operation companies that account for a majority of market shares. For privacy, we call them operator F and operator H. They are both station-based car-sharing operators. The differences between them are the number of electric vehicles and stations, vehicle models, and charging mode. We will consider the potential demand heat combined with the existing order heat to give suggestions on site location.

### 2. Data

We divide Chengdu into 58724 squared grids each of which has an area of 0.5km×0.5km. All the data we use are allocated to these grids. The data sources consist of 3 parts. The first one is the order data. We collect the order data from these two car-sharing operator mobile apps and compute the average daily order numbers for each car-sharing point. The time period is between March 28, 2018, and April 17, 2018. One order is only considered once at the beginning point. For each grid, we sum up the day-average order numbers of all points according to these two car-sharing operation companies in that grid. There are 1834 grids with car-sharing stations. The summary of these grids is shown in Table 1.

The second data source is the Point of Interest (POI) information of Chengdu that is collected from AMAP in July 2018. AMAP is an e-map that is similar to Google map and is widely used in China. However, the coordinate system of AMAP is called “GCJ-02” which is different from “WGS-84” coordinate system used in Google map. So a coordinate transformation work is introduced to ensure that the POI information is based on “WGS-84” coordinate system. The total number of POIs reaches 860196 and they are categorized into fourteen classes shown in the first column of Table 2. In AMAP POI categories, auto service contains various services such as filling station, auto-mobile rental, and charging station. But here we only consider one subcategory: auto-mobile rental since car-rental service will affect the demand of car-sharing service. Additionally, we separate transportation service into five parts: Bus Station, Underground Station, Train Station, Airport, and Parking Lots. Again we allocate these POIs to these grids and the summary of them is shown in Table 2.

The last data source used in our research is the population data (POP) called Gridded Population of the World (GPW) from Socioeconomic Data and Applications Center in NASA. Now it is the fourth version that models the population counts and densities on a continuous global raster surface. The population data is collected from the population and housing censuses between 2005 and 2014, which is used to estimate the population for the year 2000, 2005, 2010, 2015, and 2020. Some adjustments of a set of estimates based on national level, historic, and future, population predictions from the United Nation’s World Population Prospects report.
Table 2: Summary of POI information in grids.

<table>
<thead>
<tr>
<th>Category</th>
<th>Abbreviation</th>
<th>Type</th>
<th>Mean</th>
<th>Std.Error</th>
<th>Max</th>
<th>Min</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auto Service</td>
<td>CRS</td>
<td>Numerical</td>
<td>0.0221</td>
<td>0.1836</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>Food &amp; Beverages</td>
<td>FB</td>
<td>Numerical</td>
<td>2.4100</td>
<td>12.5445</td>
<td>373</td>
<td>0</td>
</tr>
<tr>
<td>Shopping</td>
<td>SH</td>
<td>Numerical</td>
<td>4.9470</td>
<td>26.9275</td>
<td>900</td>
<td>0</td>
</tr>
<tr>
<td>Daily Life Service</td>
<td>DS</td>
<td>Numerical</td>
<td>2.4800</td>
<td>11.7907</td>
<td>226</td>
<td>0</td>
</tr>
<tr>
<td>Sports &amp; Recreation</td>
<td>SR</td>
<td>Numerical</td>
<td>0.3965</td>
<td>1.9526</td>
<td>78</td>
<td>0</td>
</tr>
<tr>
<td>Medical Service</td>
<td>MS</td>
<td>Numerical</td>
<td>0.6307</td>
<td>3.3228</td>
<td>148</td>
<td>0</td>
</tr>
<tr>
<td>Accommodation Service</td>
<td>AS</td>
<td>Numerical</td>
<td>0.3428</td>
<td>0.4368</td>
<td>29</td>
<td>0</td>
</tr>
<tr>
<td>Commercial House</td>
<td>CH</td>
<td>Numerical</td>
<td>0.4638</td>
<td>2.4126</td>
<td>63</td>
<td>0</td>
</tr>
<tr>
<td>Governmental Organization &amp; Social Group</td>
<td>GS</td>
<td>Numerical</td>
<td>0.4416</td>
<td>2.1701</td>
<td>78</td>
<td>0</td>
</tr>
<tr>
<td>Science/Culture &amp; Education Service</td>
<td>SS</td>
<td>Numerical</td>
<td>0.5213</td>
<td>2.7930</td>
<td>117</td>
<td>0</td>
</tr>
<tr>
<td>Transportation Service</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bus Station</td>
<td>BS</td>
<td>Numerical</td>
<td>0.2033</td>
<td>0.6535</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>Underground Station</td>
<td>US</td>
<td>Numerical</td>
<td>0.0053</td>
<td>0.0727</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Train Station</td>
<td>TS</td>
<td>Numerical</td>
<td>0.0013</td>
<td>0.0359</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Airport</td>
<td>AP</td>
<td>Numerical</td>
<td>0.0003</td>
<td>0.0184</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Parking Lots</td>
<td>PL</td>
<td>Numerical</td>
<td>0.6332</td>
<td>2.9611</td>
<td>81</td>
<td>0</td>
</tr>
<tr>
<td>Finance &amp; Insurance Service</td>
<td>FS</td>
<td>Numerical</td>
<td>0.0874</td>
<td>0.5926</td>
<td>27</td>
<td>0</td>
</tr>
<tr>
<td>Enterprises</td>
<td>EN</td>
<td>Numerical</td>
<td>1.2750</td>
<td>7.1559</td>
<td>358</td>
<td>0</td>
</tr>
</tbody>
</table>

(2015 Revision) are also introduced to these sets of years. GPW is gridded with an output resolution of 30 arc-seconds which is approximately equal to 1 km at the equator. The value for each grid is not the population number in it but a scale that reflects the level of population. In our research we use the population density in 2015 and adjust the rasters to ours as well as the population density of them [16].

3. Methodology

The true value of the potential demand is unknown since it is an insubstantial concept which is difficult to measure. One possible way is to do a full sample questionnaire but this requires high human and financial resources and the results might be biased since questionnaire contains much subjectivity. Another way is to use usage intensity as a proxy of demand [3] or use the number of bookings while this amount can only reflect the present demand that may be restricted by the number of car-sharing stations and vehicles. Therefore using a specific amount to represent the potential demand will cause deviation. An alternative way that is applied in our research is just to distinguish whether the demand of car-sharing in a grid exists or not. Therefore, the question becomes a classification problem where classification algorithms can be implemented. So far 1834 grids contain at least one car-sharing station. The distribution of order numbers in them is shown in Figure 1(a). Clearly, those with larger order numbers can be treated as high demand. Those with tiny order numbers might not actually exist as demand but there is no way to distinguish this.}

Figure 1: The distribution of average daily order numbers. (a) The histogram of average daily order numbers. (b) The cumulative probability distribution of average daily order numbers.
order numbers are reasonably defined as no demand since very small order number reflects an occasionally demand that operators do not need to satisfy with cost of renting parking space. Therefore, we choose the lower 20% and upper 20% grids as the sample set (as shown in Figure 1(b)) and let respond equal “1” for demand and “0” for no demand. What is more, we also would like to know the level of demand in a grid which is the probability that the grid belongs to the class “1”. Classification algorithms such as k-nearest neighbour (KNN) and tree models cannot provide us such probability so we choose the following four methods: Logistic Regression, Naive Bayes, Linear Discriminant Analysis, and Quadratic Discriminant Analysis.

3.1. Logistic Regression with LASSO. Logistic Regression is used to find the relationship between variables \( X = (X_1, X_2, \ldots, X_p) \) and binary response \( Y \). It models the probability that \( Y \) belongs to a particular class. If we use a linear regression model

\[
p(X) = Pr(Y = 1 | X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p
\]

(1)

to represent the probability, we might get a result that is smaller than 0 or larger than 1, which does not make sense for probability. Therefore, in order to get an output between 0 and 1, we use the logistic function

\[
p(X) = \frac{e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}}
\]

(2)

Then we have

\[
\frac{p(X)}{1 - p(X)} = e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}
\]

(3)

By taking the logarithm of both sides, we have

\[
\log \left( \frac{p(X)}{1 - p(X)} \right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p
\]

(4)

We can see that the Logistic Regression model has a log-odds which is linear in \( X \).

The coefficients of Logistic Regression models are usually estimated by maximum likelihood method. The likelihood function for \( N \) observations is

\[
L(\beta) = \prod_{i=1}^{N} p(X_i) \prod_{i', y_{i'}=0}^{N} p(X_{i'})
\]

(5)

where

\[
\beta = (\beta_0, \beta_1, \ldots, \beta_p)
\]

(6)

\[
X_i = (X_{i1}, X_{i2}, \ldots, X_{ip})
\]

(6)

The log-likelihood can be written as

\[
I(\beta)
\]

(7)

\[
= \sum_{i=1}^{N} \left[ y_i \log p(X_i; \beta) + (1 - y_i) \log (1 - p(X_i; \beta)) \right]
\]

\[
= \sum_{i=1}^{N} \left[ y_i \beta^T X_i - \log (1 + e^{\beta^T X_i}) \right]
\]

Set its derivatives with respect to \( \beta \) to zero and we can get the maximum likelihood estimators for \( \beta \). [17]

The LASSO is a shrinkage method that constrains the coefficient estimates, which can significantly reduce the variance of them. A penalized term \( \lambda \sum_{j=1}^{p} | \beta_j | \) is added to the model, where \( \lambda \geq 0 \) is a tuning parameter.

Therefore, for Logistic Regression with LASSO, we would maximize the penalized log-likelihood function:

\[
\max_{\beta} \left\{ \sum_{i=1}^{N} \left[ y_i \beta^T X_i - \log (1 + e^{\beta^T X_i}) \right] - \lambda \sum_{j=1}^{p} | \beta_j | \right\}
\]

(8)

The selecting of a good value of \( \lambda \) is critical since it significantly affects the coefficients. Cross-validation method is applied to choose the best \( \lambda \) that produces the smallest cross-validation error [18].

3.2. Naive Bayes. Assume that there are \( K \) classes and we wish to classify an observation \( X \) into one of them. According to the Bayes’ theorem,

\[
Pr(Y = k | X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}
\]

(9)

where \( Pr(Y = k | X = x) \) denotes the probability that \( X \) belongs to the class \( k; \pi_k \) is the prior probability that a random observation is from class \( k; f_k(x) = Pr(X = x | Y = k) \) is the probability density function of \( X \) which belongs to class \( k \).

For an observation \( X = (x_1, x_2, \ldots, x_p) \), Naive Bayes assumes that each feature is independent given a class \( k \). So we have

\[
f_k(x) = \prod_{j=1}^{p} f_{kj}(x_j)
\]

(10)

where \( f_{kj}(x_j) \) is the probability density function of the \( j \)-th feature given class \( k \). Then the probability of \( X \) coming from the class \( k \) is

\[
Pr(Y = k | X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}
\]

(11)

See [18].

3.3. Linear Discriminant Analysis. Linear Discriminant Analysis is based on Bayes’ theorem and assumes that the observation \( X = (x_1, x_2, \ldots, x_p) \) is generated from a multivariate Gaussian distribution with a unique class mean vector and a common covariance matrix. We write \( X \sim N(\mu_k, \Sigma) \) where \( \mu_k \) is the mean vector for the class \( k \) and \( \Sigma \) is the covariance matrix that is the same for all classes. Then the probability density function (pdf) of \( X \) that is from class \( k \) is

\[
f_k(x) = Pr(X = x | Y = k)
\]

(12)

\[
= \frac{1}{(2\pi)^{p/2} | \Sigma |^{1/2}} \exp \left( -\frac{1}{2} (x - \mu_k)^T \Sigma^{-1} (x - \mu_k) \right)
\]
The parameters of the multivariate Gaussian distributions are unknown and we need to estimate them from training data by

\[ \hat{\pi}_k = \frac{N_k}{N} \]

\[ \hat{\mu}_k = \frac{\sum_{g_i=k} x_i}{N_k} \]

\[ \hat{\Sigma} = \frac{\sum_{k=1}^{K} \sum_{g_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T}{(N - K)} \]

where \( N_k \) is the number of observations in class \( k \) and \( g_i \) is the label of \( i \)-th observation [17].

### 3.4. Quadratic Discriminant Analysis

**Quadratic Discriminant Analysis** is a bit different from Linear Discriminant Analysis, which assumes that the covariance matrices for each class are different. Then the observation \( X \) from class \( k \) has the distribution of \( N(\mu_k, \Sigma_k) \). So the pdf becomes

\[ f_k(x) = \frac{1}{(2\pi)^{p/2} |\Sigma_k|^{1/2}} \exp \left( -\frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) \right) \]

where

\[ \lambda \] equals 0.01973 that minimizes the model deviance. Seven variables, Food&Beverages, Sports&Recreation, Governmental Organization & Social Group, Bus Station, Train Station, Airport, and Parking Lots, given nonzero coefficients. Comparing to the Logistic Regression results, Food&Beverages, Governmental

### 3.4. Quadratic Discriminant Analysis

**Quadratic Discriminant Analysis** is a bit different from Linear Discriminant Analysis, which assumes that the covariance matrices for each class are different. Then the observation \( X \) from class \( k \) has the distribution of \( N(\mu_k, \Sigma_k) \). So the pdf becomes

\[ f_k(x) = \frac{1}{(2\pi)^{p/2} |\Sigma_k|^{1/2}} \exp \left( -\frac{1}{2} (x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) \right) \]

where

\[ \lambda \] equals 0.01973 that minimizes the model deviance. Seven variables, Food&Beverages, Sports&Recreation, Governmental Organization & Social Group, Bus Station, Train Station, Airport, and Parking Lots, given nonzero coefficients. Comparing to the Logistic Regression results, Food&Beverages, Governmental

### 4. Results and Discussion

As mentioned in the Section 3, we take the lower 20% and upper 20% grids as the sample set. The quantiles are 0.79 and 6.33, respectively, which means that the grids with average daily order numbers that are less than 0.79 and larger than 6.33 are chosen as the samples. The sample size is 737 of which 370 samples are of no demand and 367 are of high demand. We randomly choose 500 samples (approximately 65% of total samples) as the training set and the remaining as the test set. Since population density and POI information are in different scale, all variables are normalized by their mean and standard error (shown in Table 3).

#### 4.1. Logistic Regression

The results of Logistic Regression are shown in Table 4. Only five variables, Food& Beverages, Medical Service, Governmental Organization & Social Group, Bus Station, and Parking Lots, are significant under 5% significance level. The population factor is not significant. Food& Beverages positively affects the probability of existence of demand, while Medical Service and Governmental Organization & Social Group has negative effects. According to our search and investigation, the car-sharing stations in Governmental Organization are always of limited access which means only staffs that work in them can get access to the stations. Once a car is returned at these stations, it always takes a long time to receive a next order. The negative effect of Medical Service could be attributed to the limited parking space and huge parking demand. The stations might be occupied by fuel vehicles. The coefficient of Bus Station is also positive, which is the same as the results in literatures [3, 19]. Parking Lots has a positive impact on the probability of existence of demand, which implies that more space provide more opportunities to build car-sharing stations. However, Chen's paper [3] concludes that more parking space leads to more private vehicle trips instead of car-sharing service.

#### 4.2. Logistic Regression with LASSO

Logistic Regression with LASSO introduces a penalized term to the model. The results are shown in Table 5. Since LASSO is a shrinkage method that the coefficients of nonsignificant variables will shrink to zero, there is no p-value for this method and p-value also does not make sense for biased regressions such as LASSO. \( \lambda \) equals 0.01973 that minimizes the model deviance. Seven variables, Food & Beverages, Sports & Recreation, Governmental Organization & Social Group, Bus Station, Train Station, Airport, and Parking Lots, give nonzero coefficients. Comparing to the Logistic Regression results, Food & Beverages, Governmental

### Table 3: Mean and standard error of all predictors.

<table>
<thead>
<tr>
<th>CRS</th>
<th>FB</th>
<th>SH</th>
<th>DS</th>
<th>SR</th>
<th>MS</th>
<th>AS</th>
<th>TA</th>
<th>CH</th>
<th>GS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.3</td>
<td>32.35</td>
<td>57.49</td>
<td>29.31</td>
<td>4.93</td>
<td>6.32</td>
<td>5.83</td>
<td>0.37</td>
<td>5.12</td>
</tr>
<tr>
<td>Std.error</td>
<td>0.67</td>
<td>43.51</td>
<td>90.96</td>
<td>34.67</td>
<td>6.83</td>
<td>8.32</td>
<td>14.78</td>
<td>1.24</td>
<td>7.51</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SS</th>
<th>BS</th>
<th>US</th>
<th>TS</th>
<th>AP</th>
<th>PL</th>
<th>FS</th>
<th>EN</th>
<th>POP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>7.34</td>
<td>1.51</td>
<td>0.072</td>
<td>0.008</td>
<td>0.004</td>
<td>8.89</td>
<td>1.23</td>
<td>16.24</td>
</tr>
<tr>
<td>Std.error</td>
<td>11.35</td>
<td>1.4</td>
<td>0.26</td>
<td>0.09</td>
<td>0.06</td>
<td>10.09</td>
<td>2.23</td>
<td>33.81</td>
</tr>
</tbody>
</table>

### Table 4: The results of Logistic Regression.

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Std.error</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.2569</td>
<td>0.1211</td>
</tr>
<tr>
<td>FB</td>
<td>0.9960</td>
<td>0.2404</td>
</tr>
<tr>
<td>MS</td>
<td>-0.3401</td>
<td>0.1581</td>
</tr>
<tr>
<td>GS</td>
<td>-0.5480</td>
<td>0.1383</td>
</tr>
<tr>
<td>BS</td>
<td>0.2814</td>
<td>0.1307</td>
</tr>
<tr>
<td>PL</td>
<td>1.3981</td>
<td>0.2182</td>
</tr>
</tbody>
</table>
Table 5: The results of Logistic Regression with LASSO.

<table>
<thead>
<tr>
<th></th>
<th>Intercept</th>
<th>FB</th>
<th>SR</th>
<th>GS</th>
<th>BS</th>
<th>TS</th>
<th>AP</th>
<th>PL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coefficient</td>
<td>0.1318</td>
<td>0.4652</td>
<td>0.1587</td>
<td>-0.2717</td>
<td>0.1641</td>
<td>0.01721</td>
<td>0.0336</td>
<td>1.0114</td>
</tr>
</tbody>
</table>

Table 6: Mean vectors for classes: demand=0 and demand=1.

<table>
<thead>
<tr>
<th></th>
<th>CRS</th>
<th>FB</th>
<th>SH</th>
<th>DS</th>
<th>SR</th>
<th>MS</th>
<th>AS</th>
<th>TA</th>
<th>CH</th>
<th>GS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Class 0</td>
<td>-0.2109</td>
<td>-0.4108</td>
<td>-0.2850</td>
<td>-0.3978</td>
<td>-0.3692</td>
<td>-0.2867</td>
<td>-0.2561</td>
<td>-0.1173</td>
<td>-0.3205</td>
<td>-0.0866</td>
</tr>
<tr>
<td>Class 1</td>
<td>0.1552</td>
<td>0.3963</td>
<td>0.3203</td>
<td>0.4089</td>
<td>0.3401</td>
<td>0.2754</td>
<td>0.2615</td>
<td>0.0390</td>
<td>0.3431</td>
<td>0.1573</td>
</tr>
<tr>
<td></td>
<td>SS</td>
<td>BS</td>
<td>US</td>
<td>TS</td>
<td>AP</td>
<td>PL</td>
<td>FS</td>
<td>EN</td>
<td>POP</td>
<td></td>
</tr>
<tr>
<td>Class 0</td>
<td>-0.3451</td>
<td>-0.2901</td>
<td>-0.0778</td>
<td>-0.0905</td>
<td>-0.0014</td>
<td>-0.4602</td>
<td>-0.2755</td>
<td>-0.2301</td>
<td>-0.1000</td>
<td></td>
</tr>
<tr>
<td>Class 1</td>
<td>0.3720</td>
<td>0.2864</td>
<td>0.0947</td>
<td>0.0435</td>
<td>0.0622</td>
<td>0.5125</td>
<td>0.3065</td>
<td>0.3012</td>
<td>0.1076</td>
<td></td>
</tr>
</tbody>
</table>

Table 7: Mean and standard deviation for the predictors grouped by class.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean Class 0</th>
<th>Std.error Class 0</th>
<th>Mean Class 1</th>
<th>Std.error Class 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRS</td>
<td>-0.2109</td>
<td>0.7259</td>
<td>0.1552</td>
<td>1.1169</td>
</tr>
<tr>
<td>FB</td>
<td>-0.4108</td>
<td>0.6177</td>
<td>0.3963</td>
<td>1.1262</td>
</tr>
<tr>
<td>SH</td>
<td>-0.2850</td>
<td>0.6372</td>
<td>0.3203</td>
<td>1.2895</td>
</tr>
<tr>
<td>DS</td>
<td>-0.3978</td>
<td>0.6668</td>
<td>0.4089</td>
<td>1.0306</td>
</tr>
<tr>
<td>SR</td>
<td>-0.3692</td>
<td>0.5560</td>
<td>0.3401</td>
<td>1.0806</td>
</tr>
<tr>
<td>MS</td>
<td>-0.2867</td>
<td>0.8778</td>
<td>0.2754</td>
<td>1.0337</td>
</tr>
<tr>
<td>AS</td>
<td>-0.2561</td>
<td>0.3561</td>
<td>0.2615</td>
<td>1.3175</td>
</tr>
<tr>
<td>TA</td>
<td>-0.1173</td>
<td>0.5521</td>
<td>0.0390</td>
<td>0.6665</td>
</tr>
<tr>
<td>CH</td>
<td>-0.3205</td>
<td>0.7296</td>
<td>0.3431</td>
<td>1.1695</td>
</tr>
<tr>
<td>GS</td>
<td>-0.0866</td>
<td>0.9128</td>
<td>0.1573</td>
<td>1.1696</td>
</tr>
<tr>
<td>SS</td>
<td>-0.3451</td>
<td>0.4995</td>
<td>0.3720</td>
<td>1.2910</td>
</tr>
<tr>
<td>BS</td>
<td>-0.2901</td>
<td>0.8360</td>
<td>0.2864</td>
<td>1.0089</td>
</tr>
<tr>
<td>US</td>
<td>-0.0778</td>
<td>0.8589</td>
<td>0.0947</td>
<td>1.1439</td>
</tr>
<tr>
<td>TS</td>
<td>-0.0905</td>
<td>0.0000</td>
<td>0.0435</td>
<td>1.2157</td>
</tr>
<tr>
<td>AP</td>
<td>-0.0014</td>
<td>0.9907</td>
<td>0.0622</td>
<td>1.4038</td>
</tr>
<tr>
<td>PL</td>
<td>-0.4602</td>
<td>0.6131</td>
<td>0.3125</td>
<td>1.1715</td>
</tr>
<tr>
<td>FS</td>
<td>-0.2754</td>
<td>0.6450</td>
<td>0.3065</td>
<td>1.2305</td>
</tr>
<tr>
<td>EN</td>
<td>-0.2301</td>
<td>0.5698</td>
<td>0.3012</td>
<td>1.3990</td>
</tr>
<tr>
<td>POP</td>
<td>-0.1000</td>
<td>0.7764</td>
<td>0.1076</td>
<td>1.1253</td>
</tr>
</tbody>
</table>

Organization & Social Group, Bus Station, and Parking Lots have the same effect. Medical Service is no longer significant in this model; instead Sports & Recreation positively affects the probability of existence of demand. The reason could be that Sports & Recreation is mostly visited by young people who have high acceptance and usage intensity of car-sharing mode [20]. In this model Train Station and Airport are also significant which could be attributed to their functions of traffic hub that brings high exposure rate and large mobile population. The population factor is again not significant.

4.3. Linear Discriminant Analysis. As mentioned in Section 3.3, Linear Discriminant Analysis (LDA) assumes that the sample $x_i$ is from a multivariate Gaussian distribution where the mean vector is unique for different class and the covariance matrix is the same. Table 6 shows the mean vectors for classes: demand = 0 and demand = 1. The covariance matrix is shown in Appendix A. The prior for these two classes are $\pi_0 = 0.502$ and $\pi_1 = 0.498$.

4.4. Quadratic Discriminant Analysis. The mean vector of Quadratic Discriminant Analysis (QDA) is the same of LDA. However, QDA assumes that the covariance matrix is different for these two classes. The results are shown in Appendix B.

4.5. Naive Bayes. Naive Bayes assumes that each predictor is independent and can be represented by a Gaussian distribution within each group. The mean and standard deviation for other variables grouped by class are shown in Table 7. The Train Station is a constant within class: demand = 0 since
Table 8: AUC value and accuracy rate for the previous five models.

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC value</th>
<th>Accuracy Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>0.8500</td>
<td>0.7722</td>
</tr>
<tr>
<td>Logistic regression with LASSO</td>
<td>0.8545</td>
<td>0.7637</td>
</tr>
<tr>
<td>LDA</td>
<td>0.8513</td>
<td>0.7553</td>
</tr>
<tr>
<td>QDA</td>
<td>0.8020</td>
<td>0.6835</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>0.8146</td>
<td>0.7215</td>
</tr>
</tbody>
</table>

4.6. Comparing These Five Models. To judge the performance of a classifier, AUC value and accuracy rate are always applied. AUC value is the area under receiver operating characteristic (ROC) curve that varies between 0 and 1. The higher this value, the better the discrimination will be. Accuracy rate is computed by the number of correct predictions divided by the total number of predictions. We test these models with the remaining 237 observations. The values of these two measures for the previous five models are shown in Table 8. We can see that all models give an AUC value between 0.8 and 0.9 and accuracy rate between 0.65 and 0.8. QDA model produces the least AUC value and accuracy rate while the pure Logistic Regression or Logistic Regression with LASSO performs best since their AUC value and accuracy rate are the largest. LDA model and Naive Bayes model are a bit worse than Logistic Regression. Therefore, we can choose either the pure Logistic Regression or Logistic Regression with LASSO as the final model. Here we choose the Logistic Regression with LASSO model.

4.7. Optimizing Location of Car-Sharing Stations. For all of the 58724 squared grids, we normalize them with the mean and standard error in Table 3 and then run the Logistic Regression with LASSO model. The predictive probabilities of existence of demand for these grids against present order numbers are shown in Figure 2. The left figure (a) shows the results for the whole 58724 squared grids. We can see that most grids are with order numbers less than 10 so to be clear we plot these grids in figure (b). We define predictive probability larger than 0.5 as high demand heat and otherwise low demand heat. What is more, it is reasonable to consider average daily order numbers less than 1 as low order heat and otherwise high order heat. Therefore, the grids can be sorted in to 4 groups as shown in Figure 2(b): I: high demand heat and high order heat; II: high demand heat and low order heat; III: low demand heat and low order heat; IV: low demand heat and high order heat. For stations in group III grids, operators are advised to close or remove them after investigation such as checking the time interval between two orders. For stations in group IV, further work is required to check if other influential factors that are omitted here exist since low demand heat and high order heat mutually conflict. We will mainly focus on groups I and II when optimizing the locations of car-sharing stations. Figure 3 shows the grids with high demand heat in Chengdu. The red colour refers to group I and the blue colour refers to group II. It is obvious that these grids concentrate in the city centre and town centre that is also the gathering area of crowd and business.

For the grids in groups I and II, we can give suggestions to the two operators: operator F and operator H on the optimal location of car-sharing stations. Three cases of grids are considered:

(i) Case 1: grids with no operator F stations and at least one operator H stations
Figure 3: Grids with high demand heat in Chengdu: red for grids with high order heat; blue for grids with low order heat.

Figure 4: Case 1: grids with no operator F stations and at least one operator H stations.

(ii) Case 2: grids with no operator H stations and at least one operator F stations

(iii) Case 3: grids with no operator H stations and no operator F stations

Figures 4–6 display these three cases. In Case 1, we can see that operator H occupies most of the space in city centre and a majority of grids are of high order heat which means the operation effect is quite good for operator H in these grids. Only several grids are of low order heat where operator F can consider building stations. In Case 2, hundreds of grids of high demand heat surrounding the city centre are monopolized by operator F. However in the north-west of Chengdu, most grids are of low order heat, which suggests an opportunity for operator H to build stations in them. Moreover, in Case 3, we find that these two operators have not entered town centres yet where most of these grids are of high demand heat. These two operators are advised to do investigation at these areas and consider building stations.

5. Conclusion

This research focuses on optimizing the car-sharing stations in Chengdu market. The main methodology is trying to estimate the potential demand combining with the present order numbers. Unlike the previous research that applies multiple linear regression to model the demand, this study transforms the question to a binary problem whether the demand exists or not. Then five classification models are...
introduced to model this and estimate the probability of existence of demand. Three data sources, average daily order numbers from operator F and operator H, POI information, and population data, are used. From the five models we summarize as follows:

(i) In the Logistic Regression, Food& Beverages, Bus Station, and Parking Lots have positive effect on the probability of existence of demand, while Medical Service and Governmental Organization & Social Group have negative effect.

(ii) The Logistic Regression with LASSO model indicates that Food& Beverages, Sports& Recreation, Bus Station, Train Station, Airport, and Parking Lots positively affect the probability of existence of demand and Governmental Organization & Social Group have opposite effect. This model result contains the traffic hubs that could increase the demand of car-sharing service.

(iii) Linear Discriminant Analysis (LDA) model and Quadratic Discriminant Analysis (QDA) model compute the prior probability that is 0.502 for class: demand = 0 and 0.498 for class: demand = 1. The mean vector and covariance matrix of multinomial distribution are then estimated to compute the posterior.

(iv) In the Naive Bayes model, each variable is assumed to be normally distributed of which the mean and standard deviation are estimated with each class. The standard deviation of Train Station in class:demand = 0 is zero; thus it is treated as a constant.

(v) Comparing the performance of these five models by AUC value and accuracy rate, we find that QDA models give the worst estimation while Logistic Regression and Logistic Regression with LASSO perform best. LDA is a bit worse than these two models and Naive Bayes works slightly better than QDA. Therefore we conclude that linear models work better in our case.

The Logistic Regression with LASSO is chosen as the final model and used to estimate the probability of existence of demand for all grids. The predictive probability larger than 0.5 is treated as high demand heat and otherwise low demand heat. These grids with high demand heat are concentrated in the city centre or town centre. Together with the present order numbers, 4 groups of grids are defined. Different suggestions on optimizing location of car-sharing stations are given for each group. Both operator F and operator H are advised to build stations in the absent grids with high demand heat and close or remove part of stations in the grids with low demand heat and low order heat. Operator H is also advised to build stations in the north-west of Chengdu where high demand exists and operator F has low operation efficiency.

However, there are still some limitations in our research.

(i) First, Chengdu is a competitive market with more than five car-sharing operators of which only two are considered in our research. The suggestions may be reliable without consideration of the other operators.

(ii) Second, our research is based on 500m∗500m squared grids which are suitable in city centre but are too small for other areas since the building density is much lower in these areas.

(iii) Third, one important factor, cost of building stations, is not considered in our search. The cost always includes renting the parking space, building charging piles, and purchase electric vehicle. The definition of high order heat and low order heat in our research is simply based on the order number equalling one. However, high cost stations require high order numbers to redeem the cost. Therefore, such definition should vary with the cost of building stations.

(iv) Fourth, samples are chosen based on the lower 20% and upper 20% grids according to average daily order numbers subjectively. Other approaches can be applied such as lower 30% and upper 30% to include more observations. The definition of 4 groups is also subjective. More objective method such as clustering may be applied.

(v) Fifth, the models we use are all based on strong assumptions. More technical classification algorithm can be applied such as bagging, boosting, random forest, and Gaussian Process.

(vi) Sixth, the order numbers only consider the one that is placed and omit the return behaviour. One station may have low pick-up orders while having high drop-off orders which is called usage imbalance. The decision on such station should be made carefully.

(vii) Seventh, variable selection work can be done before we train the classification model since not all predictors are related to the response.

Furthermore, it is worth investigating the effect of adopting those suggestions by estimating the order numbers when a certain number of stations are added. One possible way is to first find the features of relationship between start and end grids and compute the intensity. Then for a grid with some new added stations, possible related grids can be found and the order number can also be estimated by intensity. Another possible way is to do transportation simulation where all of the transportation modes, the population, and the traffic operation are considered, which is a huge and challenging work.

**Appendix**

Covariance matrices for LDA and QDA and distributions of predictors for Naive Bayes.

A. Covariance Matrix for LDA

Covariance matrix for LDA
### Covariance matrix of class demand\(=0\) for QDA

The covariance matrix for class demand\(=0\) for QDA is given by:

\[
\begin{bmatrix}
    0.1912 & 0.0741 & 0.2556 & 0.1233 & 0.2016 & 0.0143 \\
    0.1031 & 0.3169 & 0.1434 & 0.0767 & 0.1995 & 0.0010 \\
    0.0356 & 0.1491 & 0.0721 & 0.0383 & 0.0786 & 0.0003 \\
    0.3837 & 0.2432 & 0.1866 & 0.1338 & 0.0817 & 0.0002 \\
    0.2140 & 0.1386 & 0.0839 & 0.0545 & 0.0741 & 0.0001 \\
    0.7246 & 0.4759 & 0.3472 & 0.2314 & 0.1543 & 0.0000 \\
\end{bmatrix}
\]

### B. Covariance Matrix for QDA

The covariance matrix of class demand\(=0\) for QDA is represented as:

\[
\begin{bmatrix}
    0.1912 & 0.0741 & 0.2556 & 0.1233 & 0.2016 & 0.0143 \\
    0.1031 & 0.3169 & 0.1434 & 0.0767 & 0.1995 & 0.0010 \\
    0.0356 & 0.1491 & 0.0721 & 0.0383 & 0.0786 & 0.0003 \\
    0.3837 & 0.2432 & 0.1866 & 0.1338 & 0.0817 & 0.0002 \\
    0.2140 & 0.1386 & 0.0839 & 0.0545 & 0.0741 & 0.0001 \\
    0.7246 & 0.4759 & 0.3472 & 0.2314 & 0.1543 & 0.0000 \\
\end{bmatrix}
\]

### Covariance matrix of class demand\(=1\) for QDA

The covariance matrix for class demand\(=1\) for QDA is given by:

\[
\begin{bmatrix}
    0.1912 & 0.0741 & 0.2556 & 0.1233 & 0.2016 & 0.0143 \\
    0.1031 & 0.3169 & 0.1434 & 0.0767 & 0.1995 & 0.0010 \\
    0.0356 & 0.1491 & 0.0721 & 0.0383 & 0.0786 & 0.0003 \\
    0.3837 & 0.2432 & 0.1866 & 0.1338 & 0.0817 & 0.0002 \\
    0.2140 & 0.1386 & 0.0839 & 0.0545 & 0.0741 & 0.0001 \\
    0.7246 & 0.4759 & 0.3472 & 0.2314 & 0.1543 & 0.0000 \\
\end{bmatrix}
\]
Figure 7: Continued.
C. Distributions of Predictors for Naive Bayes

See Figure 7.

Data Availability

All data included in this study are available upon request by contact with the author Yu Cheng (chengyu@shevdc.org). They will all be based on the grids defined in the research since order data for a specific station is quite sensitive and private.

Disclosure

This research was presented at the conference "International Conference on Smart Mobility and Logistics in Future Cities" and presentation slides which include brief introduction of this research are shared to that conference. This manuscript is also published in the main website of our organization for internal study and communication.

Conflicts of Interest

The authors declared that they have no conflicts of interest to this work.

Acknowledgments

This work is supported by International Science & Technology Cooperation Program of China under Contract no. 2016YFE0102200.
References


